Прежде всего сформилируем условия задачи:

1. Выбрать ген - выбран **Ribosomal protein S10**. Далее выбрать виды, в геноме которых есть данный ген. Минимальное число видов, необходимых для включения в дерево – 10. Большее число видов 15-20 приветствуется. Найти аминокислотные последовательности генов для нижеперечисленных видов и составить fasta-файл с последовательностями. Поставить в начало названия гена строку с названием организма, чтобы организм отображался на листьях дерева. Например:

   > Bacterium AHI16583.1 ATP synthase alpha [Lactobacillus sp. HRR0T3]

Хорошо бы иметь представленность следующих видов:

Человек - human

Обезьяна - primates

Грызуны – mouse, rat

копытное - bovine

сумчатое - marsupials

пресмыкающиеся – snakes, lizards, turtles

птица

рыба

растение

fungi (Saccharomyces cerevisiae) – пекарские дрожжи (продаются сухими в пакетиках)

архея - archaea

бактерия – bacteria

Поиск гена производится на сайте NCBI в базе данных “protein”. Набиваете название гена в скобках. (ribosomal protein L1) + используйте кнопку Advanced для лучшего поиска.

2. Произвести множественное выравнивание последовательностей алгоритмами ClustalW и Muscle из пакета Mega. Сохранить два выравнивания в отдельные файлы. Написать, есть ли разница в выравнивании.
3. В программе Mega построить филогенетическое дерево для аминокислотных последовательностей с бутсрэп-анализом для двух выравниваний – и методом ClustalW, и Muscle.

- методом расстояний (UPGMA)

    - методом расстояний (NJ)
- методом максимального правдоподобия

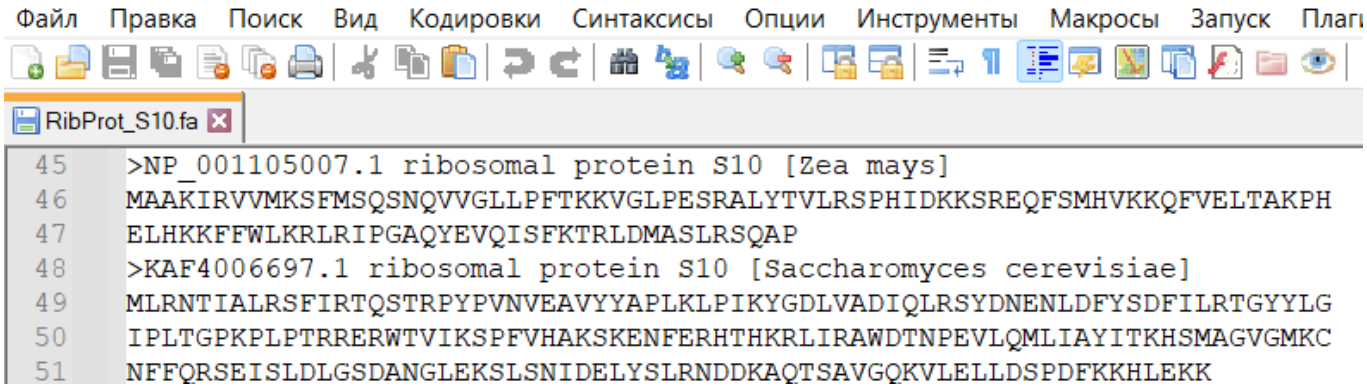
В качестве отчетности представить файл в формате fasta с аминокислотными последовательностями файлы с двумя выравниваниями – clustalw и muscle отчет, оформленный в виде pdf или ptt: скриншот выравнивания для метода ClustalW и Muscle скриншоты деревьев, построенных (1) методом расстояний (UPGMA), (2) методом расстояний (NJ), методами максимального правдоподобия, на которых будут видны бутстрэп-значения, а также хорошо читаемые названия организмов. Выводы Какой алгоритм выравнивания лучше сработал - ClustalW или Muscle? Одинаковая ли получилась топология деревьев при построении разными методами? Одинаковые ли получились бутстрэп-значения? Совпадают ли деревья, построенные по одному гену с принятыми деревьями видов?

Скачивание данных:

Для начала выберем представителей для каждой из запрашиваемых категорий:

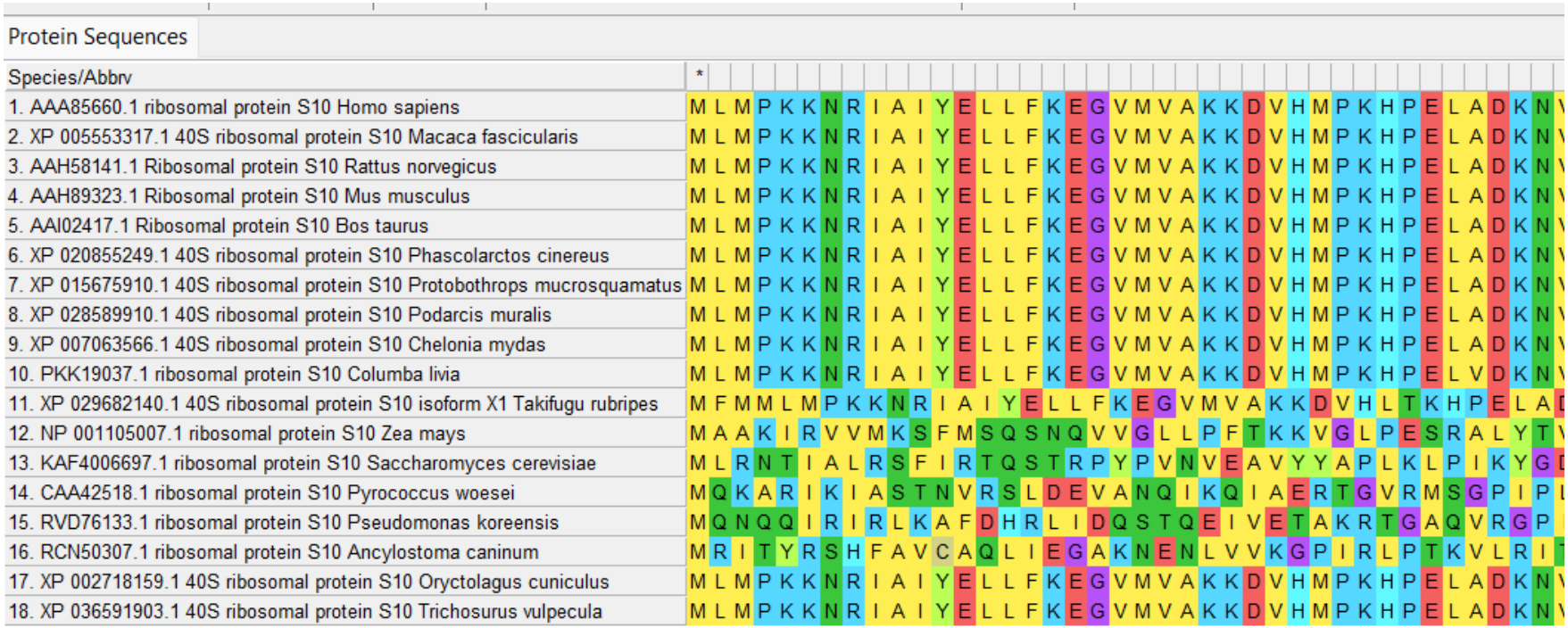
1. Человек - Homo sapiens
2. Обезьяна - Macaca fascicularis
3. Крыса - Rattus norvegicus
4. Мышь - Mus musculus
5. Копытное - Bos taurus
6. Сумчатое - Phascolarctos cinereus
7. Змея - Protobothrops mucrosquamatus
8. Ящерица - Podarcis muralis
9. Черепаха - Chelonia mydas
10. Птица - Columba livia
11. Рыба - Takifugu rubripes
12. Растение - Zea mays
13. Пекарские дрожжи - Saccharomyces cerevisiae
14. Архея - Pyrococcus woesei
15. Бактерия - Pseudomonas koreensis
16. Ракообразные - Ancylostoma caninum
17. Кролик - Oryctolagus cuniculus
18. Кускусовые - Trichosurus vulpecula

После чего полученные последовательности объединяем в один .fa файл:



Выравнивание:

Для начала рассмотрим результат:



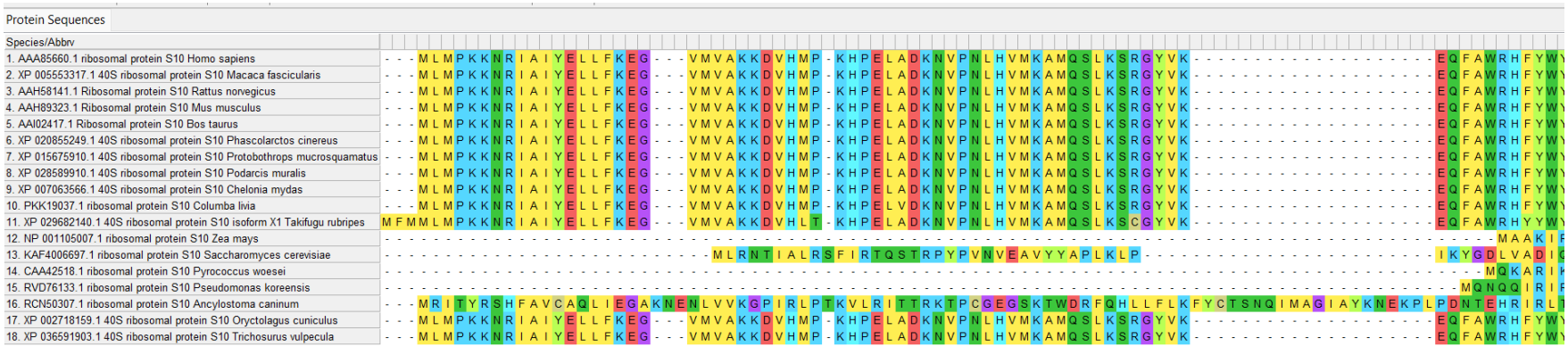
Видно, что для разных организмов последовательность одинаковая, что говорит о консервативности гена.

ССЫЛКА НА ДАННЫЕ:

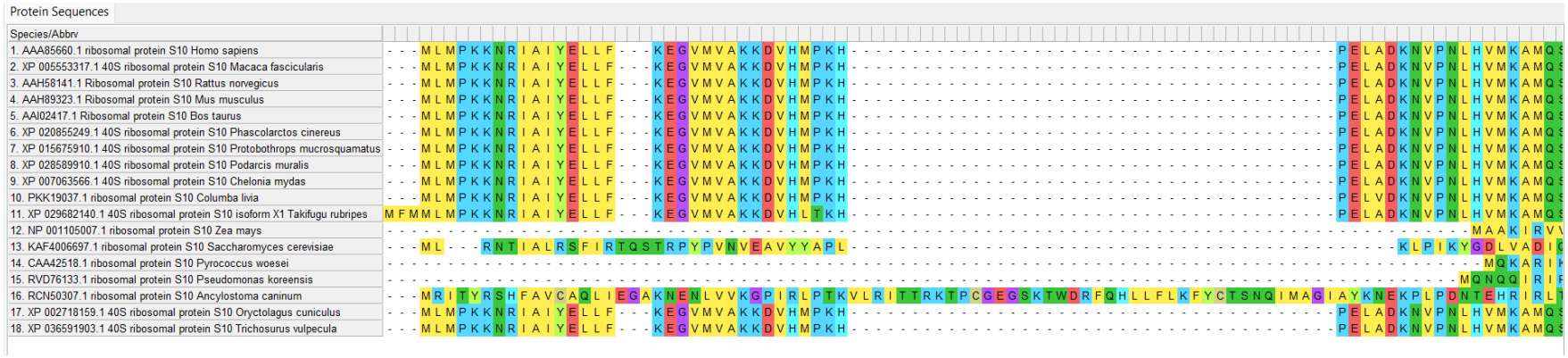
https://drive.google.com/drive/folders/1xblFq3xWgH81rSToDC_TXKGB96DIZTnc?usp=*sharing*

Теперь сделаем выравнивание двумя способами.

ClustalW:



MUSCLE:



Можно видеть, что получившиеся выравнивания действительно отличаются. Можно видеть, что в целом, все последовательности кроме 12-16 почти идентичны (различия видны только дальше).

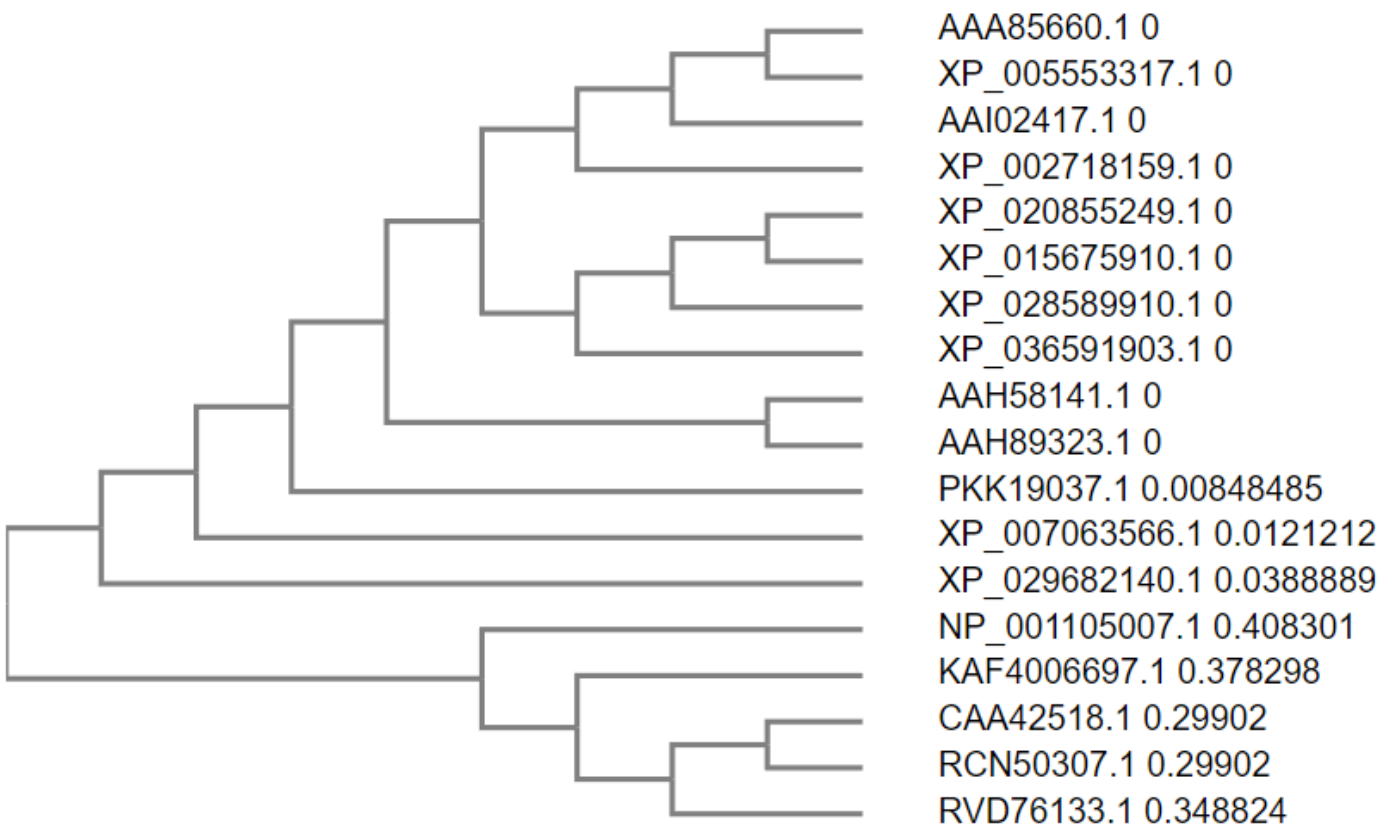
Построение деревьев:

Теперь для каждого выравнивания будем строить выравнивания одним из 3 способов. Перед построением дерева нажимаем Phylogenetic analysis (!), только после этого выбираем bootstrap (пусть со значением 100) и строим дерево по выбранному методу.

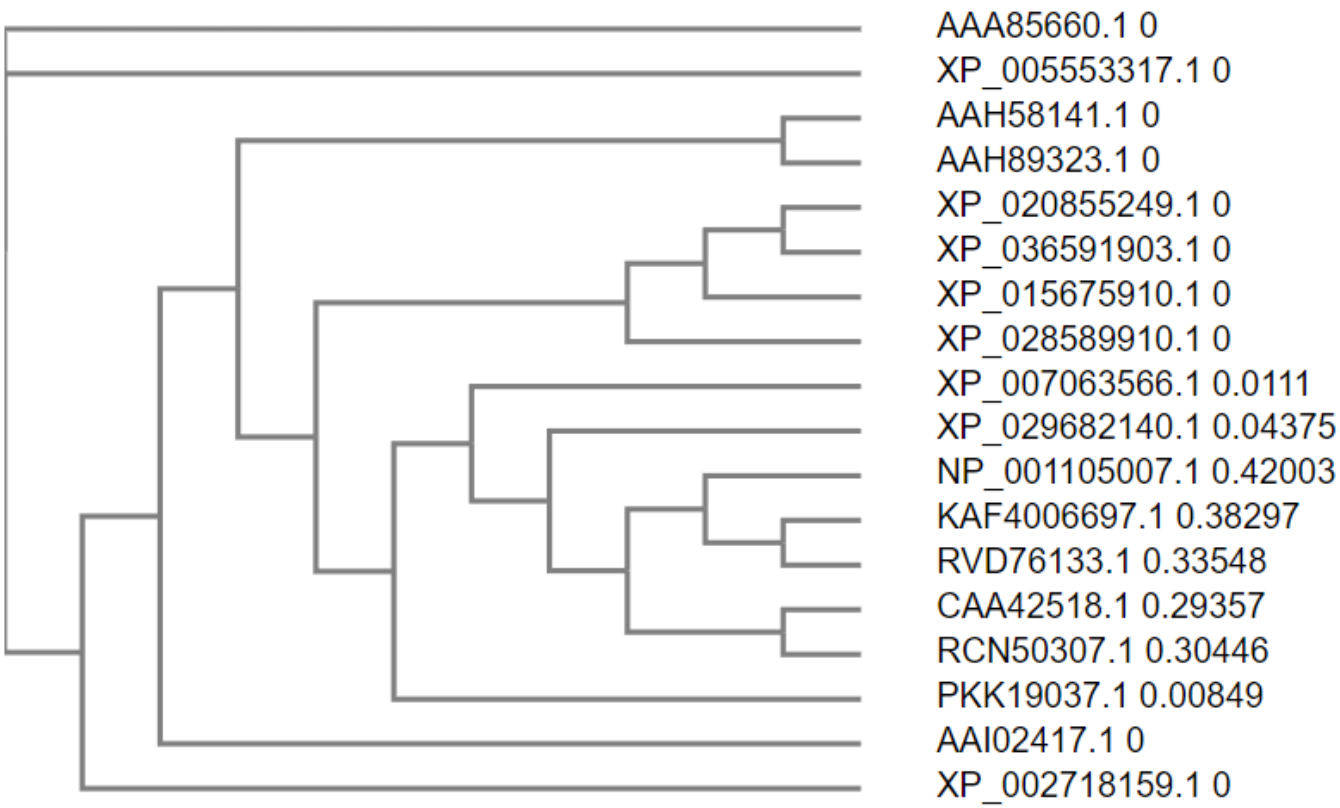
ClustalW:

MEGA вылетает для UPGMA и NJ, обсуждал это все с Коноваловым, он сказал делать на стороннем сайте, но там нет бутстрепа, но тем не менее:

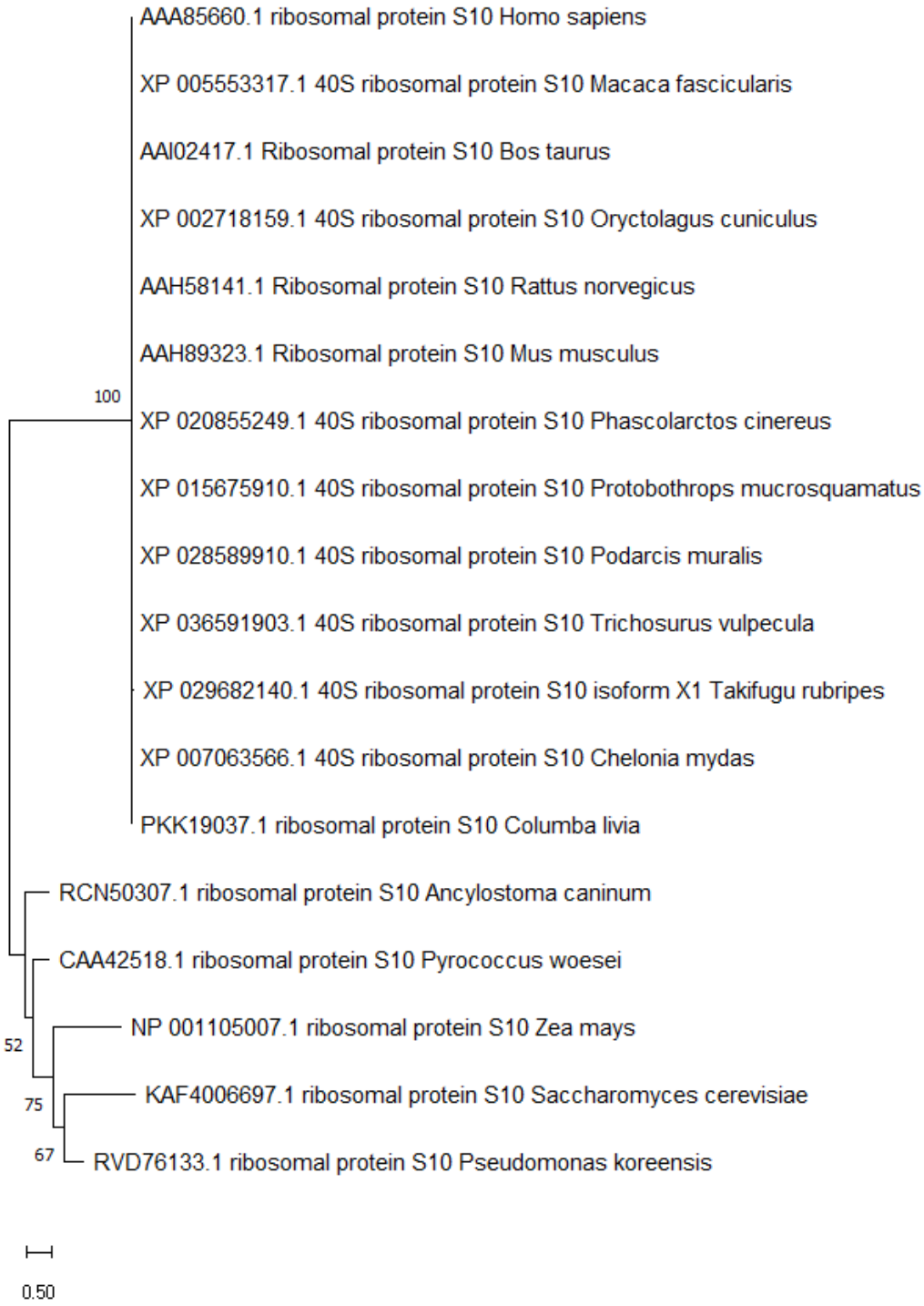
UPGMA:



Neighbor joining:

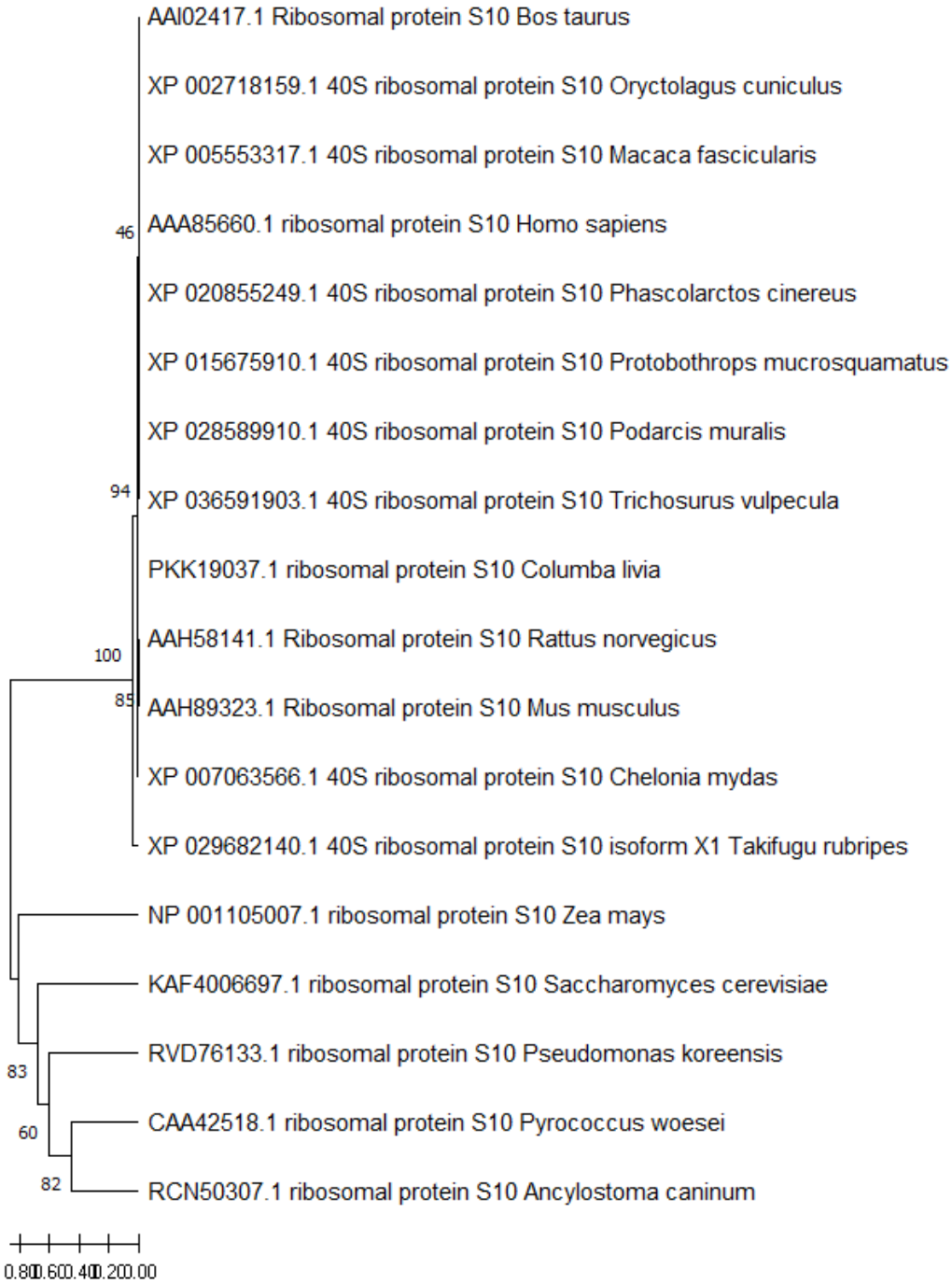


Maximum likelihood:

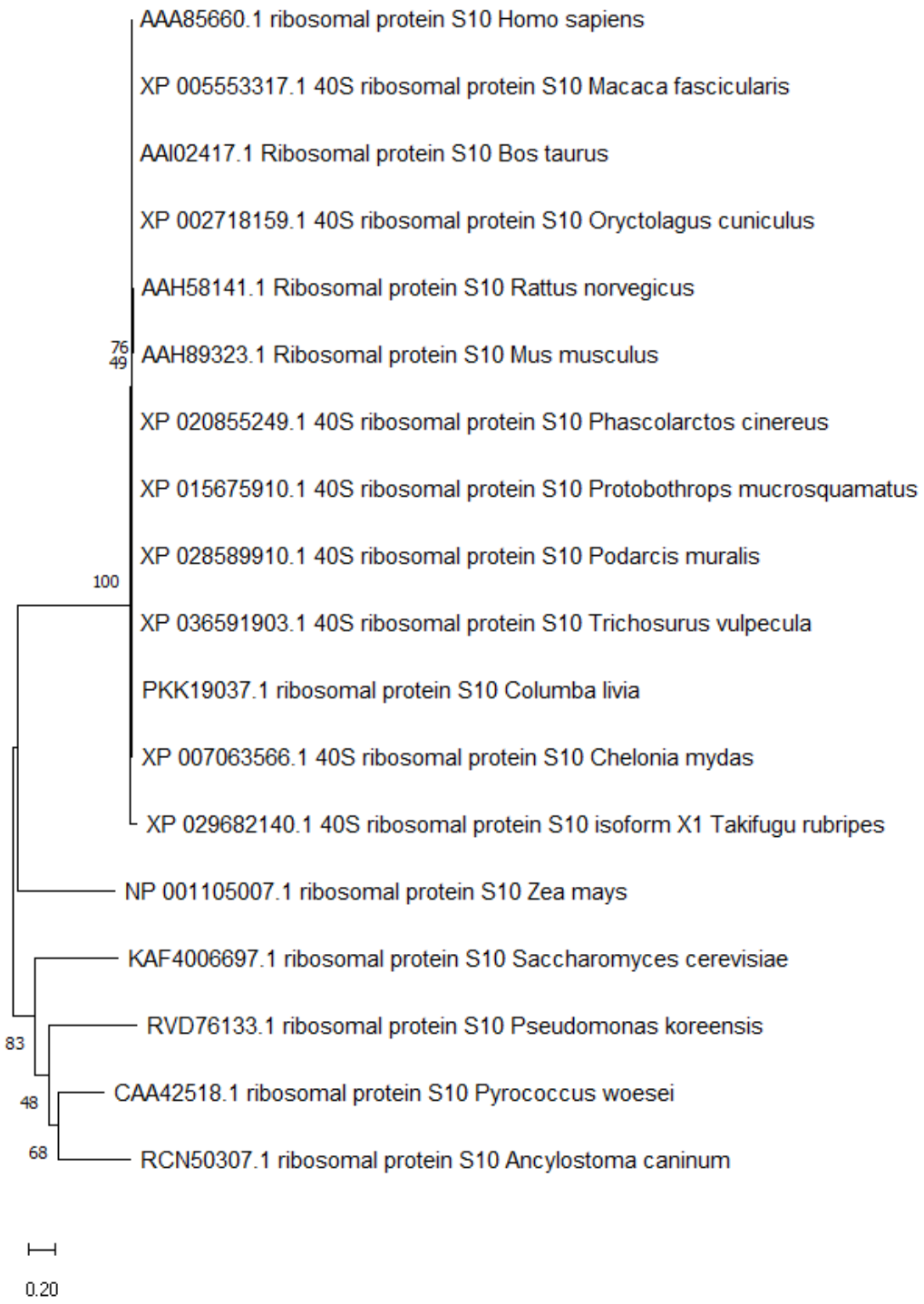

AAA85660.1 ribosomal protein S10 Homo sapiens
XP 005553317.1 40S ribosomal protein S10 Macaca fascicularis
AAI02417.1 Ribosomal protein S10 Bos taurus
XP 002718159.1 40S ribosomal protein S10 Oryctolagus cuniculus
AAH58141.1 Ribosomal protein S10 Rattus norvegicus
AAH89323.1 Ribosomal protein S10 Mus musculus
100
XP 020855249.1 40S ribosomal protein S10 Phascolarctos cinereus
XP 015675910.1 40S ribosomal protein S10 Protobothrops mucrosquamatus
XP 028589910.1 40S ribosomal protein S10 Podarcis muralis
XP 036591903.1 40S ribosomal protein S10 Trichosurus vulpecula
XP 029682140.1 40S ribosomal protein S10 isoform X1 Takifugu rubripes
XP 007063566.1 40S ribosomal protein S10 Chelonia mydas
PKK19037.1 ribosomal protein S10 Columba livia
RCN50307.1 ribosomal protein S10 Ancylostoma caninum
CAA42518.1 ribosomal protein S10 Pyrococcus woesei
NP 001105007.1 ribosomal protein S10 Zea mays
52
KAF4006697.1 ribosomal protein S10 Saccharomyces cerevisiae
75
67
RVD76133.1 ribosomal protein S10 Pseudomonas koreensis
0.50

MUSCLE:

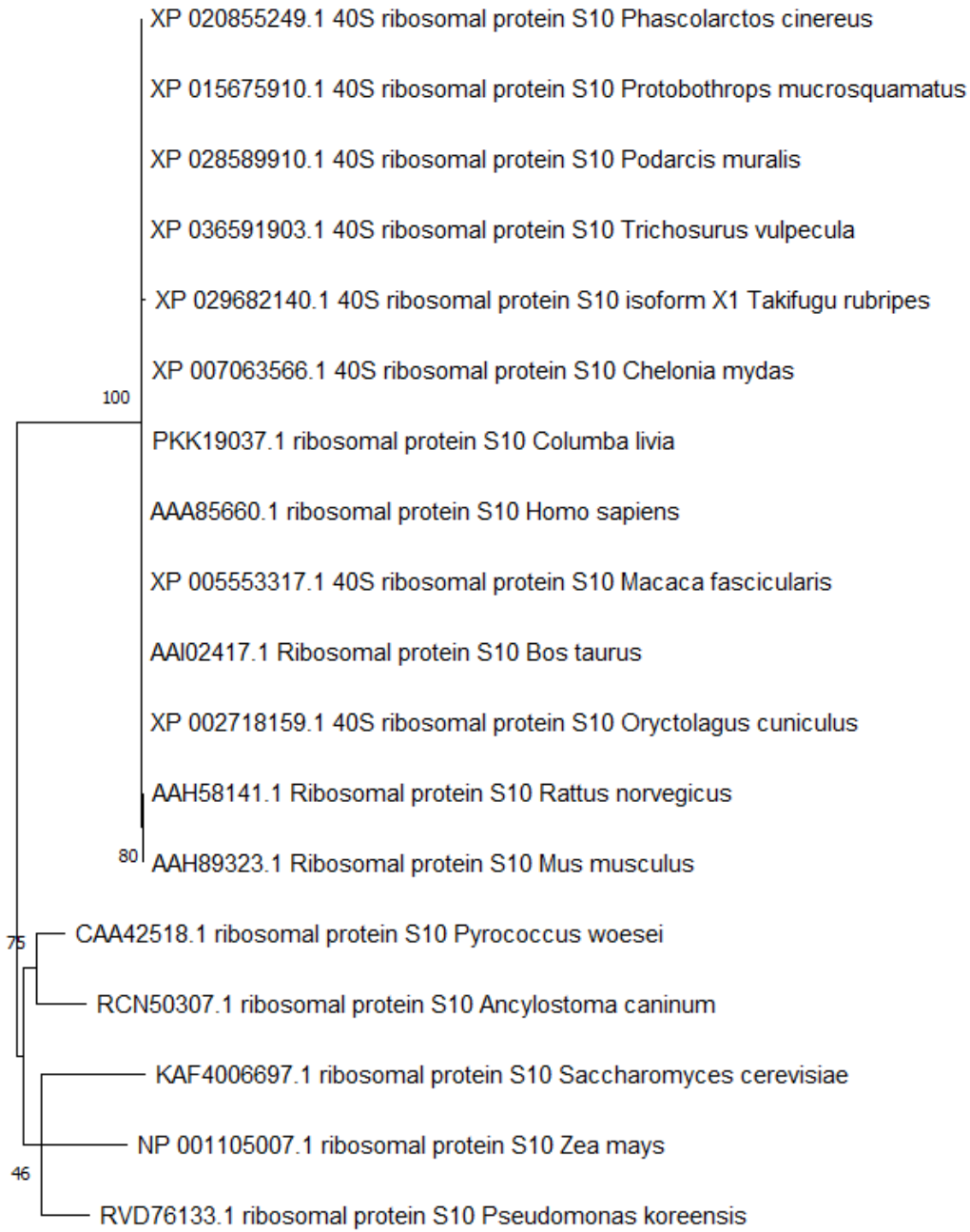
UPGMA:


AAI02417.1 Ribosomal protein S10 Bos taurus
XP 002718159.1 40S ribosomal protein S10 Oryctolagus cuniculus
XP 005553317.1 40S ribosomal protein S10 Macaca fascicularis
46
AAA85660.1 ribosomal protein S10 Homo sapiens
XP 020855249.1 40S ribosomal protein S10 Phascolarctos cinereus
XP 015675910.1 40S ribosomal protein S10 Protobothrops mucrosquamatus
XP 028589910.1 40S ribosomal protein S10 Podarcis muralis
94
XP 036591903.1 40S ribosomal protein S10 Trichosurus vulpecula
PKK19037.1 ribosomal protein S10 Columba livia
AAH58141.1 Ribosomal protein S10 Rattus norvegicus
100
85
AAH89323.1 Ribosomal protein S10 Mus musculus
XP 007063566.1 40S ribosomal protein S10 Chelonia mydas
XP 029682140.1 40S ribosomal protein S10 isoform X1 Takifugu rubripes
NP 001105007.1 ribosomal protein S10 Zea mays
KAF4006697.1 ribosomal protein S10 Saccharomyces cerevisiae
RVD76133.1 ribosomal protein S10 Pseudomonas koreensis
83
60
CAA42518.1 ribosomal protein S10 Pyrococcus woesei
82
RCN50307.1 ribosomal protein S10 Ancylostoma caninum
0.80 0.60 0.40 0.20 0.00

Neighbor joining:



Maximum likelihood:



Ответы на вопросы:

1. Какой алгоритм выравнивания лучше сработал - ClustalW или Muscle?

Ответ: Если смотреть на все выравненные последовательности для каждого метода, то Muscle получается сильно лучше. Он лучше "определил" начало и конец последовательностей, идейно целостней выглядит результат. То есть в случае с ClustalW результат выглядит так:

Начало - пропуск - кусок - пропуск... -кусок - пропуск,

то для Muscle видно и начало, и конец, то есть:

Начало - пропуск - кусок - пропуск... -кусок - пропуск - **конец**, поэтому я считаю, что он справился лучше. (**да и устроен он тоже лучше, как алгоритм и идея**)

2. Одинаковая ли получилась топология деревьев при построении разными методами?

Ответ: Мне немного трудно воспринимать результаты из другого приложения (для ClustalW использовали), поэтому давайте рассмотрим на примере Maximum likelihood метода.

Глобально говоря, для обоих выравниваний получилось две одинаковых по набору группы, но если смотреть более подробно, то можно видеть, что "объединялись" в группы разные последовательности, поэтому именно одинаковой назвать нельзя, но они весьма похожи.

И заметил странную деталь, что для MUSCLE выравнивания, в самом низу дерева последовательности KAF4006697.1 И RVD76133.1 как будто бы не пренадлежат дереву, что очень странно, хотя, возможно, это просто баг.

3. Одинаковые ли получились бутстрэп-значения?

Ответ: Опять же, судя по Maximum likelihood методу, это не так. Это наблюдается в нижней половине дерева.

4. Совпадают ли деревья, построенные по одному гену с принятыми деревьями видов?

Ответ: Думаю, что не совсем справедливо судить по данному дереву, поскольку в верхней половине последовательности генов абсолютно идентичны (вроде бы даже все). Но, если отвечать на вопрос, то нет, не стоит, поскольку на одном уровне находится человек, мышь, копытное, птица, кролик. А, например с бактериями и археями находится ракообразное..